The CHIL RT07 Evaluation Data

Susanne Burger

interACT, Carnegie Mellon University
407 South Craig Street, Pittsburgh 15213, USA
sburger@cs.cmu.edu
http://www.is.cs.cmu.edu

Abstract. This paper describes the CHIL 2007 evaluation data set provided for the Rich Transcription 2007 Meeting Recognition Evaluation (RT07) in terms of recording setup, scenario, speaker demagogic and transcription process. The corpus consists of 25 interactive seminars recorded at five different recording sites in Europe and the United States in multi-sensory smart rooms. We compare speakers' talk-time ratios in the interactive seminars with lecture data and multi-party meeting data. We show that the length of individual speaker's contributions helps to position interactive seminars between lectures and meetings in terms of speaker interactivity. We also study the differences between the manual transcription of narrow-field and far-field audio recording.
Keywords: multi-modal, data collection, transcription, meetings, interactive seminars

1 Introduction

For several years researchers have been interested in different aspects of how participants in multi-party meetings interact with each other. This has continuously led to the creation of large-scaled research programs, projects and international evaluations of technologies around this topic. One of the recent projects is CHIL - Computers in the Human Interaction Loop [1], an Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme. CHIL started in January 2004 and will finish its work in August 2007. 15 partners from nine countries are jointly coordinated by the Universität Karlsruhe (TH), Germany and the Fraunhofer Institute IITB, Germany. Based on the understanding of human perception, CHIL computers are enabled to provide helpful assistance implicitly, requiring minimal human attention or interruptions. To serve development and evaluation of the CHIL technologies, multi-sensory audiovisual lecture and seminar data was recorded inside smart rooms (CHIL rooms) at five different CHIL partner sites located in Europe and the United States.

In 2005 CHIL partners started to participate in NIST's rich transcription (RT) [2] and multi-modal evaluations such as CLEAR [3]. NIST extended their test and training data sets from multi-party meetings (conference room scenario) to lecture type data (lecture room scenario) to accommodate new evaluations such as speaker activity detection and source localization. The basic differences

between lecture and conference room data are the number and setting of meeting participants, their interactivity and the addition of far-field microphone arrays and extensive usage of video in the lecture data collection.

CHIL contributed a development and test data set to the Rich Transcription 2007 Meeting Recognition Evaluation (RT07) which consists of 25 seminars recorded in 2006 [4], five seminars per recording site. These seminars are supposedly more interactive than the lecture room data CHIL contributed to previous evaluations. This paper describes in a brief summary the technical setup and the situation in which participants were recorded. The following sections report on speaker demagogies, transcription and how test set segments were picked from the recorded episodes. Finally, we compare the ratio of speaker's talk-time in the interactive seminars with the ratios of other lecture- and meeting-type data and show the differences between far-field and narrow-field transcription.

2 The CHIL Rooms

The CHIL 2006 seminars [4] were recorded in smart seminar rooms, called CHIL rooms (see also figure 1). These are seminar rooms which provide multiple recording sensors, audio as well as video. There are five different recording sites with completely equipped CHIL rooms:

- AIT: Research and Education Society in Information Technologies at Athens Information Technology, Athens, Greece
- IBM: IBM T.J. Watson Research Center, Yorktown Heights, USA
- ITC-irst: Centro per la ricerca scientifica e technologica at the Instituto Trentino di Cultura, Trento, Italy
- UKA: Interactive Systems Labs of the Universität Karlsruhe, Germany
- UPC: Universitat Politècnica de Catalunya, Barcelona, Spain

Having different sites in multiple countries benefited the variability in the collected data due to the different sizes of the rooms, layouts and light features. In particular, the site variability supplied a range of European English such as British English, American English and and a range of English with foreign accents from all over the world. To ensure a homogeneous technical recording quality, each site equipped its room with a minimum base-set of conformed and identical hardware and software.

2.1 Sensor Setup

The minimum video equipment required in a CHIL room includes four fixed corner cameras, a panoramic camera on the table and at least one pan-tilt-zoom (PTZ) camera.

The minimum setup for audio recording comprises far-field and narrow-field sensors. The far-field (FF) data is collected through at least one NIST Mark III microphone array (developed by NIST) [5], which consists of 64 small microphones in a row, and is mounted on the smart room's wall. The Mark III channels

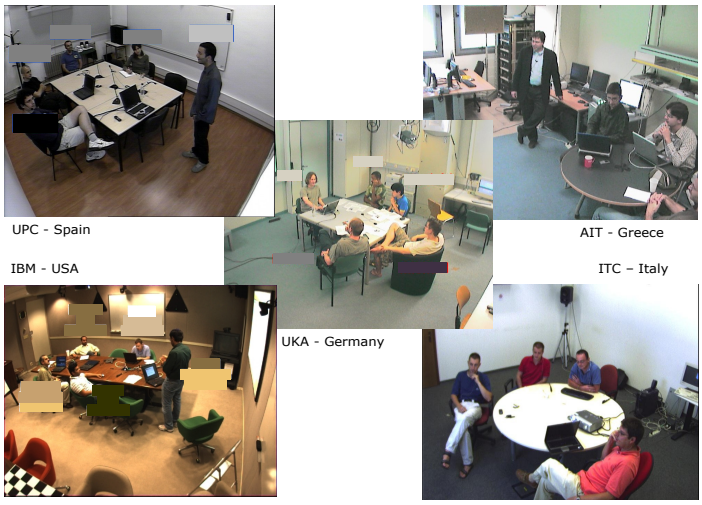

Fig. 1. Single-camera views recorded at the five CHIL rooms during interactive seminars

are recorded in SPHERE format via an Ethernet connection to a recording computer in the form of multiplexed IP packages.

A minimum of three T-shaped four-channel microphone arrays are mounted on the room's walls. At least three table top microphones are placed on the meeting table, distributed an appropriate distance from each other.

The narrow-field audio data is collected through close-talking microphones (CTM). The presenter wears a wireless microphone because presenters tend to stay standing and move around more frequently. The basic setup for the other participants consists of one close-talking microphone per participant, wireless if possible. In comparison, the CHIL lectures recorded in previous years had contributions from speakers from the audience which were only picked up by far-field microphones.

The T-shaped arrays, the table top microphones and all CTMs are amplified by RME Octamic 8 channel amplifiers and are recorded via Hammerfall HDSP9652 I/O sound cards. All audio recordings, including the Mark III array channels, were sampled in 44 khz, 28 bits.

As an example for a completely equipped CHIL room, figure 2 shows a sketch of the IBM CHIL room. IBM uses two Mark III arrays, four T-shaped arrays, and three table top microphones.

3 Interactive Seminars

The seminars recorded in 2006 are *interactive seminars*: three to five participants sit around a seminar table while one person presents research work. The other

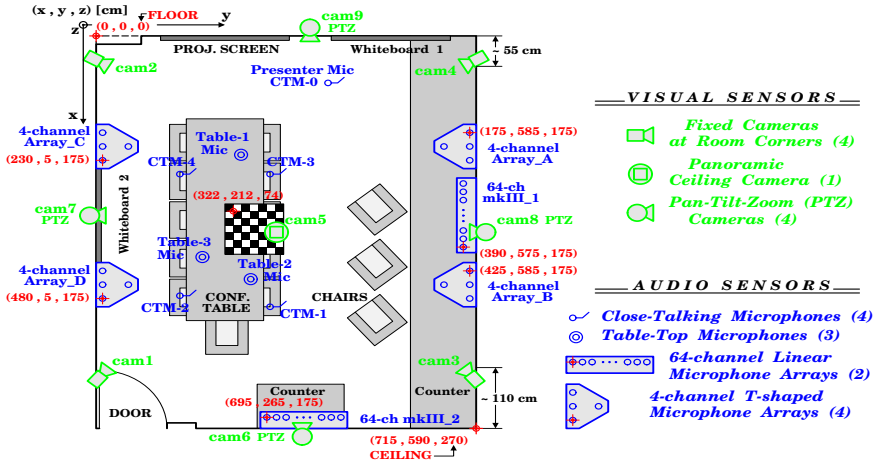

Fig. 2. Sketch of the IBM CHIL room

participants may interrupt at any time, ask questions, make comments, give suggestions. This frequently leads to real discussions and meeting-type conversation resulting in frequent interaction between the participants. On the contrary, the CHIL lectures recorded in previous years provided less opportunity for discussion or casual conversation.

The term *scripted* is often used in relation to acted scenarios, where participants' contributions are from following scripts, certain actions are predefined or where participants act in pre-given roles. The CHIL seminars are real seminars which were scheduled without the purpose of data collection. Speaker's contributions occurred naturally and spontaneously. However, to support the evaluation and the development of the multiple CHIL technologies, participants were asked to produce acoustic events from a given list, for example, door slam, chair moving, applause, laugh, cough, keyboard typing. In most of the seminars, a participant would receive a cell phone call. Individuals would come late to the seminar or leave early. There would be a short coffee break in the middle of the seminar. Not all of these extra features were recorded in each seminar. Unexpectedly, these artificially included events elicited spontaneous reactions and contributions of the participant and added humorous scenes. These in turn enriched the naturalness of the data instead of constraining it.

4 Speakers

71 individuals spoke in the CHIL seminars. Unfortunately, only five of them are female. The speakers originate from all over the world, mainly from Europe. All of them speak English, most of them with a foreign accent, with the biggest groups being Spaniards (23%), Italians (15%), and Greeks and Germans (each 14%). All sites had visitors, foreign colleagues or students participating in their

recordings so that there is a total of 17 countries represented. Figure 3 shows the distribution of the speakers' countries of origin for the CHIL RT07 test- and development set speakers.

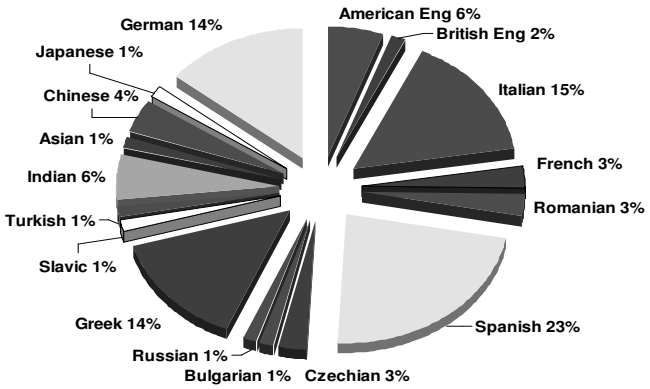

Fig. 3. Speaker accents: Distribution of countries of origin in the CHIL RT07 development-set and test-set

5 Transcription

The manual transcription of the speech in the audio recordings was done at Carnegie Mellon University. Transcribers started by transcribing all channels of the close-talking microphones on word level in ISL style (e.g. including labels for vocal noises such as laughter, coughing and filled pauses, tags for word breaks, neologisms, repetitions and corrections). The speaker contributions were manually segmented into talk spurts. In [9], talk spurts have been defined as "speech regions uninterrupted by pauses longer than 500 ms" . The CHIL reference segmentation used a minimum threshold of inter-spurt duration of 300 ms. This value has recently been adapted for the purpose of building speech activity detection references in the NIST RT evaluations.

The transcription of the far-field condition was based on one of the table-top microphone recordings, usually the most centered microphone or the table microphone channel with the best audio quality. Since a significant portion of the transcription remains the same in both conditions, transcribers adapted the narrow-field transcription to what they perceived from the far-field channel. Changes include removing speaker contributions which were not picked up by

the table microphone. Very softly spoken utterances or voiceless laughter which frequently was not audible at all in the far-field recording was also removed. Mumbled utterances or those interfering with noise or speaker overlap were substituted with the label for non-identifiable. In the cases where transcribers could barely recognize a word, they transcribed their best guess and marked the word as hard-to-identify. Contributions or details which were only audible through the table microphone were inserted in the transcription. These were the rare instances where participants had removed their close-talking microphones to leave or to get coffee. Sometimes the CTM recording was too clipped, had technical problems or interfered with another sound source and thus could not be transcribed in the narrow-field condition.

Transcribers used the annotation tool TransEdit[1], a tool developed in-house for multi-channel meeting transcription. It is easy and intuitive to use, independent of the user's education. It focuses on support and convenience for the sole purpose of the transcription and segmentation of speech. It displays all audio channels in parallel which is very helpful when listening to interactive multi-channel conversations. The parallel view also allows the comparison of audio recordings in different qualities.

TransEdit transcriptions result in two annotation files: the actual transcription and the segments' time stamps in sample point values. These files were converted and combined into the NIST STM format. The following section shows a short excerpt of a narrow-field transcription of a CHIL seminar in STM format:

```
...
ait_20060728_ctm ctm_3 ait_20060728_ctm-ait_004 269.69 273.335 <o,male>
so I should sell +/all/+ <uh> all my property in Crete ?
ait_20060728_ctm ctm_1 inter_segment_gap 269.947 273.022 <o,>
ait_20060728_ctm ctm_1 ait_20060728_ctm-ait_005 273.022 277.697 <o,male>
no . <P> build it so that it becomes the basis for knowledge-based
applications .
ait_20060728_ctm ctm_3 inter_segment_gap 273.335 276.206 <o,>
ait_20060728_ctm ctm_3 ait_20060728_ctm-ait_004 276.206 279.724 <o,male>
so% +/I will/+ I will put some *smartness in my olive trees
ait_20060728_ctm ctm_1 inter_segment_gap 277.697 279.825 <o,>
ait_20060728_ctm ctm_4 ait_20060728_ctm-ait_003 279.623 280.631 <o,male>
<Laugh>
...
```

6 Evaluation Data Selection

For evaluation purposes, the transcribed data was separated into development-set data and test-set data. The development-set contains a total of 2 hours 45 minutes of recording. It comprises five complete seminars, each recorded by one of the five CHIL room sites. The average duration of a seminar is 33 minutes. Four to five participants spoke at each seminar. The development-set includes

[1] TransEdit is available for research purposes by sending email to sburger@cs.cmu.edu

6,656 talk spurt segments, 44,300 word tokens and 2,729 unique word types. 10% of the word types are proper names.

The test-set consists of 40 seminar segments of approximately five minutes each. These segments were selected from the remaining 20 seminars; each of the five recording sites collected four of them. To provide a balanced assortment of the different sections of the seminars, the colleagues of the Universität Karlsruhe developed a system which chooses segments of

- the beginning of a seminar (including the arrival of the participants, welcoming, introduction),
- the actual talk or presentation (including other participants' questions and comments),
- the coffee break or any other section of casual conversation,
- the question and answer part or discussion part at the end of a presentation,
- the end of the seminar (including closing, planning, good-bye, departure of the participants).

Each of the categories was represented at least one time per site and in similar distribution over all sites. The total duration of the test-set is 3 hours 25 minutes. The set consists of 11,794 talk spurts, 56,196 word tokens and 2,870 unique word types. 9.5% of the word types are proper names.

The participants of the RT07 evaluation used development and test-set data of the RT04 and RT05 evaluations as training data for their systems.

7 Speaker Talk-time Ratios

The *talk-time* [6] of a speaker is the total duration of all segmented talk spurts. This includes speech pauses shorter that 300 ms and vocal noises. The ratio of the talk-time of individual speakers during a meeting describes who dominated the meeting in terms of talking for the longest period of time. It also represents how the talk-time was distributed between the participants. The talk-time is calculated as percentage of the total duration of the meeting per individual speaker. A total of all speakers' talk-time will most likely not sum up to 100% because speakers' contributions overlap frequently. The possible pauses between contributions also add to the total duration of a meeting.

To prove that the new interactive CHIL seminars have more interaction between the participants than the previous lecture recordings, we compared speakers' talk-time ratios of the CHIL development- and test-sets with other lecture and multi-party meeting corpora. We looked at six ISL lectures (a total of 6 hours 38 minutes recording), 18 ISL student presentations (8 hours 47 minutes), 27 ISL conference talks (15 hours 45 minutes), 19 meetings of the ISL meeting corpus (10 hours 13 minutes, [7] and [8]) and 28 five-minute segments of 14 of the CHIL 2004 lectures. The latter were part of the CHIL evaluation test-set of the the RT05 evaluation. The ISL lecture, presentation and talk recordings were collected at CMU between 1999 and 2005 for lecture recognition and machine translation projects.

For each corpus, we took the average of the talk-time of the speakers who talked for the longest time (group A), the speakers who talked for the second longest time (group B) and for the third longest time (group C), see figure 4. The CHIL development-set and the CHIL test-set were analyzed separately because the recording setting was slightly changed before and after the collection of the development data. In order to have more data, we looked at the complete 20 test-set seminars, which were the source of the five-minute selections for the CHIL test-set, for a total duration of 11 hours 7 minutes.

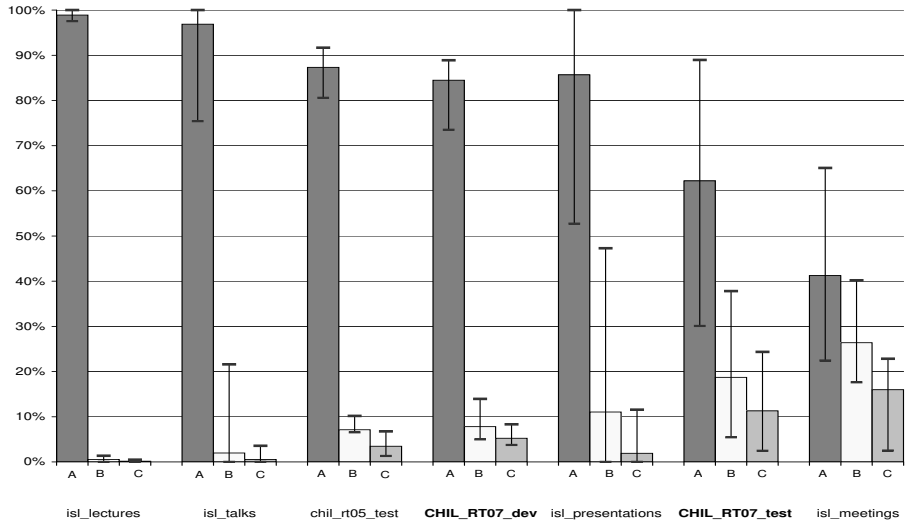

Fig. 4. Comparison of speakers' talk time ratio in different data-sets averaged for the speakers with the longest talk-time (A), second longest talk-time (B) and third longest talk-time (C) per data set. Maximum and minimum talk-time duration for each data-set are shown by the little lines on top of each bar.

Figure 4 sorts the seven data sets by duration of talk-time for speakers of the group A.

The ISL lectures and the ISL talks show talk-time durations for group A of 98% and 97%, respectively. The other participants accounted for almost none or only very short talk-time during these recordings. The lectures have a formal teaching session setting; a docent teaches a class to an audience of students. Very rarely do the students ask questions at the end of the lecture. They never interrupt. The talks are research presentations at scientific conferences and workshops. The audience only contributes at the time-for-question period at the end of the talk and also never interrupts during the talk.

The talk-time ratios for the CHIL RT05 lectures and the CHIL RT07 development data seminars look very much a-like with slightly more activity for groups B and C in the RT07 data than in the RT05 data. Participants account for more

contributions than in the ISL lectures and talks, because the setting was less formal. The presenter-audience relationship was not always a teacher-student relationship but rather frequently an adviser-student or even peer-to-peer relationship. The RT07 development data were the first recordings done in 2006. At this time, the data collection transitioned to *interactive* seminars. Thus these first recordings were not yet as interactive as the later recorded seminars.

Similar to the RT05 lectures and the RT07 development-set, the ISL presentations were student presentations with an audience of students and advisers who interrupted more frequently. There is also a significantly broader variety between the talk-time durations of groups A, B and C. This is different from all data sets before. It shows that the setting was not as formal or fixed as it was in the lectures and talks.

Each CHIL recording site collected four more seminars during the summer of 2006, the CHIL RT07 test-set. It offered a higher degree of freedom to interact and occasions for casual conversations (as described in section 3). The result for the talk-time ratios in these seminars show much shorter sections of monologue of one single speaker, in average for group A 62%, and longer periods of talk-time for groups B and C. There is also more variety of talk-time ratios of the seminars in this data-set. Some seminars come close to ratios as thus can be seen in the ratios of the ISL multi-party meetings. Here group A's talk-time duration lasts 41% on average, group B speaks for 26% of the meeting on average. Group C still speaks for 16% of the time in the meeting.

As a result, the comparison of talk-time ratios of lecture-type data and meeting-type data positions the CHIL interactive seminars in between. The seminars provide both casual conversation and discussion, as well as monologue-type presentation.

8 Far-field vs Narrow-field Transcription

The effort to prepare separate transcriptions for the far-field condition as well as for the narrow-field condition is large: the transcription of the narrow-field quality of the CHIL seminars took about 20 times real time including the second passes; the far-field transcription added another 10 times real time on average to the transcription task. To study what was actually changed during the far-field transcription pass, we compared the differences in far-field and narrow-field transcriptions and counted the added and removed elements.

Figure 6 shows the loss and the gain of transcribed elements in terms of what was removed from the narrow-field transcription and what was added. The values are displayed as loss and gain in percentage over both CHIL RT07 data-sets on average, in comparison to the number of elements in the CTM transcription. This is shown for each recording site.

Accordingly, transcribers removed from the close-talking microphone transcription an average of 4% of complete talk spurts, 2% of word tokens, 15% of word fragments (wrdfrgm) and 12% of laughter annotations. The far-field tran-

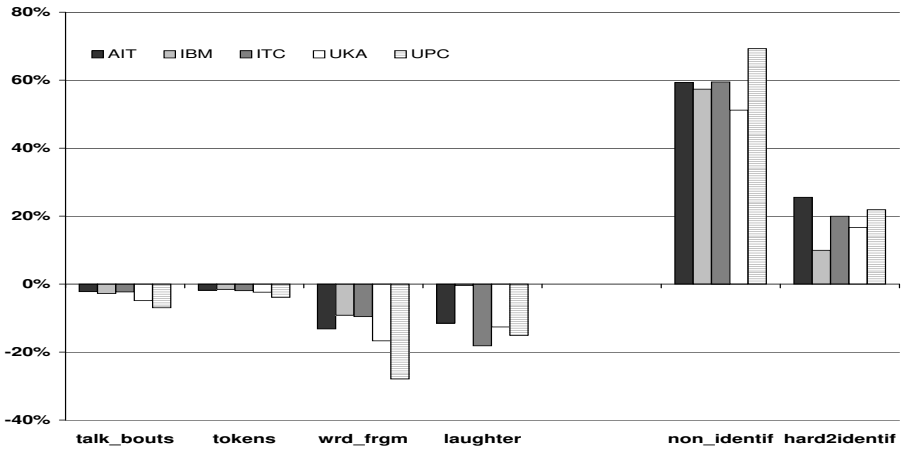

Fig. 5. Percentage of loss and gain of transcribed elements of the table microphone transcription compared to the close-talking microphone transcription in average of the CHIL RT07 development and test-set data, for each recording site.

scriptions show an average of 60% more labels for non-identifiable utterances (non-identif) and 19% more word tokens tagged as hard to identify (hard2identif).

9 Conclusion

We described the CHIL evaluation data sets for the RT07 evaluation in terms of recording sensors and setup, scenario, speaker demagogic and transcription process.

We compared speakers' talk-time ratios of the CHIL evaluation data with other lecture-type data and multi-party meeting-type data. The comparison allows us to prove that the CHIL interactive meetings are more interactive than lecture data, but less interactive than multi-party meeting data and, therefore, need to be placed between these categories.

We finally were able to display what is added and what is removed from narrow-field transcriptions when the recording quality is changed to far-field sensors.

The CHIL interactive seminars are publicly available to the community through ELRA's catalog of language resources (http://catalog.elra.info).

Acknowledgments. We would like to thank Matthew Bell, Brian Anna, Brett Nelson, JP Fridy and Freya Fridy for the transcription of the data and helpful input to this paper. The data collection presented here was funded by the European Union under the integrated project CHIL (Grant number IST-506909).

References

1. The CHIL Consortium Website: http://chil.server.de
2. The NIST RT Website: http://www.nist.gov/speech/tests/rt/
3. The CLEAR Evaluation Website: http://www.clear-evaluation.org
4. Mostefa, D., Potamianos, G., Chu, S.M., Tyagi, A., Casas, J., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., Rochet, C.: The CHIL Audiovisual Corpus for Lecture and Meeting Analysis inside Smart Rooms. accepted for Journal for Language Resources and Evaluation, Springer, Netherlands (2007)
5. The Mark III Website: http://www.nist.gov/smartspace/cmaiii.html
6. Nadia Mana, N., Burger, S., Cattoni, R., Besacier, L.: The Nespole! VoIP Corpora In Tourism And Medical Domains. Proc. Eurospeech 2003, Geneva, Switzerland (2003)
7. Burger, S., Maclaren, V., Yu, H.: The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. in Proc. ICSLP, Denver, CO, USA (2002)
8. Burger, S., Sloane, Z.: The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions. NIST Meeting Recognition Workshop 2004, NIST 2004, Montreal, Canada (2004)
9. Shirberg, E., Stolcke, A., Baron, D.: Observations on Overlap: Findings and implications for automatic processing of multi-party conversation. in Proc. Eurospeech, Aarlborg, Denmark (2001)